QUALITY ASSESSMENT OF ON-LINE VIDEOS USING METADATA

Chul-Hee Han and Jong-Seok Lee

School of Integrated Technology
Yonsei University, Korea
{hanch0232, jong-seok.lee}@yonsei.ac.kr

ABSTRACT

As video consumption becomes popular, demand for high quality of experience of consumed videos is also increasing. While online video sharing is a popular video application, quality assessment of videos in such an application is challenging due to lack of reference videos and simultaneous involvement of diverse quality factors. In this paper, we take advantage of additional information of online videos, i.e., metadata, and explore the extent to which video quality can be estimated from metadata. Subjective quality assessment using crowdsourcing is conducted, based on which metadata-based quality models are constructed. It is shown that the estimated quality scores show fairly high correlation with the subjective quality.

Index Terms— YouTube videos, quality assessment, metadata, crowdsourcing, paired comparison

1. INTRODUCTION

Nowadays, video services are widely spread and popularized via massive propagation of Internet service networks and devices. The amount of video data shared online is exponentially increasing. At the same time, demand for high quality video contents is also increasing. Thus, an appropriate quality assessment of shared videos would be helpful in order to, e.g., provide a filtering function based on visual quality or perform quality enhancement for low quality contents.

YouTube, the largest video sharing site, offers a limited filtering assortment regarding visual quality at this moment. In fact, quality assessment of videos shared online is challenging due to the following issues. First, due to the absence of the original reference videos, full-reference objective quality metrics, which requires both the original reference and test video data, cannot be exploited. This means that no-reference metrics are the only choice for objective quality assessment of the videos. While plausible full-reference metrics are available, development of reliable no-reference metrics is not sufficiently mature in comparison to the full-reference case [1]. Second, various elements of quality degradation are involved in the videos, such as unstable camera motion, blurring, blockiness due to compression, etc., which makes objective quality assessment of online videos more challenging. Third, it is difficult to effectively obtain reliable subjective quality assessment data due to the large volume of video data to be assessed.

Despite these challenges, there is also a unique opportunity in quality assessment of online videos. In this paper, we exploit another modality for quality assessment, i.e., metadata accompanied with the videos. We examine which metadata are helpful to estimate the video quality, and derive models predicting video quality from selected metadata. Although there exist researches to conduct no-reference quality assessment of web videos based on visual data (e.g., [2]), our work is the first attempt to investigate the feasibility of metadata for quality assessment, to the best of our knowledge. Moreover, in order to obtain ground truth subjective quality data, which is related to the third problem mentioned above, crowdsourcing is employed.

The rest of the paper is organized as follows. The following section explains how the video data and subjective ground truth data were collected. Then, Section 3 details how we process the data. The results are shown in Section 4. Finally, conclusion is given in Section 5.

2. VIDEO AND SUBJECTIVE DATA COLLECTION

2.1. Video and metadata database

In order to consider a wide range of video quality variations, we target user-generated contents (UGC) in our study. Via search in YouTube using popular keywords in UGC (e.g., video, fun, food, animal, nature, life, etc.), we collected a large set of videos. Non-UGC videos were excluded and, finally, we selected 50 videos. Using the YouTube API, we gathered the metadata of the videos, including the video IDs, maximum resolutions, uploaded dates, video lengths, view counts, ‘like’ counts, ‘dislike’ counts, ‘subscribe’ counts,

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the IT Consilience Creative Program (NIPA-2014-H0201-14-1001) supervised by the NIPA (National IT Industry Promotion Agency).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Table 1. Statistics of the collected subjective data.

	Group 1	Group 2	Group 1+2 (sum)
# total comparisons	2422	6049	8471
# presentations per video (average)	96.88	241.96	338.84
# survey days	21	3	24
# comparisons per day	115.33	2016.33	352.96

comment counts, IDs of the uploaders, numbers of other uploaded videos by the uploaders, etc.

2.2. Crowdsourced subjective quality assessment

In order to collect a large size of subjective quality assessment data, we employed crowdsourcing. It has been emerging as an effective solution for large scale subjective quality assessment, possibly at the expense of slight decrease of reliability but with significant reduction of costs for certain tasks [3]. Crowdsourcing suits perfectly to our case because the stimuli to be assessed are by nature consumed by crowd over the Internet. In particular, we chose the paired comparison methodology among standardized subjective quality assessment methodologies [4], where a pair of videos are presented and the winner is chosen by each subject. As mentioned in Section 1, several quality factors are involved in the online videos and, and as shown in [3], it is helpful to simplify the subjects' task in crowdsourcing. The paired comparison methodology imposes the minimum task load on the subjects and thus is effective to obtain reliable test results [5,6].

We created a web page for our quality assessment experiment. The page first briefly introduces the purpose and procedure of the experiment. Then, two randomly selected videos among 50 videos are shown to the subject. The videos were shown in a resolution of 480x360 (noted as 360p in YouTube) because this resolution is the default one when one opens a video in YouTube. When only lower resolutions are available for a video, the maximum available resolution was used. The subject had to click the play button of the each video to watch it. Only when both of the videos were played, the voting buttons were activated. The subject chose one between the two videos as having better visual quality, and then the next pair of videos were shown for comparison.

We conducted crowdsourcing with two different groups. The first group ("Group 1") was mainly composed of acquaintances to whom the instruction of the test was given carefully. Another group ("Group 2") consisted of totally unknown subjects hired in Amazon Mechanical Turk, and thus the instruction was given through the web page for this group. Table 1 summarizes the statistics of the subjective data collection for the two groups. 2422 and 6049 comparisons were conducted for the two groups, respectively, thus 8471 comparisons were collected in total. On average, each video was shown 338.84 times. It is interesting to see that, while it took 21 days to collect 2422 comparisons from Group 1, 6049 comparisons were gathered only within 3 days from Group 2, which means that the collection in Group 2 about 18 times faster than that in Group 1.

3. DATA PROCESSING

3.1. Quality score computation

While several possibilities exist for computing quality scores from comparison results, e.g., Elo rating, the Bradley-Terry model, the Thurstone's model, etc. [6,7], we adopted a simple method to define the winning rate as the quality score of a video, i.e.,

$$\text{Quality Score} = \frac{\#\ \text{wins}}{\#\ \text{presentations}}$$

Thus, the quality scores range within 0 and 1.

3.2. Metadata processing

The following basic metrics were obtained from the collected metadata.

- Maximum size: the height of the maximum resolution of a video (e.g., 240, 360, 1080, etc.).
- # day number: the number of days between the YouTube foundation date (February 14, 2005) and the uploaded date of a video. A more recent video has a larger value of the day number.
- Length: the video length in seconds.
- # view: the number of view counts of a video.
- # like: the number of 'like'.
- # dislike: the number of 'dislike'.
- # comment: the number of comments.
- # subscribe: the number of subscriptions.
- # uploaded video: the number of other videos uploaded by the uploader of a video.

Then, several other metrics were derived via transformation of the basic metrics, such as # like divided by # view, # comment divided by # subscribe, etc. The full list of the derived metrics can be seen in Table 2.

In total, we examined 19 measures obtained from metadata.

4. RESULTS

4.1. Correlation analysis

The performance of each metric derived from metadata for quality estimation is evaluated in terms of the linear

Table 2. Results of the correlation analysis between the metrics driven from metadata and subjective quality scores.

Metadata metrics	Group1	Group2	Group1+2
Maximum size	**0.6392**	**0.4641**	**0.5233**
# Day number	**0.4885**	**0.3033**	**0.3579**
# Like /# View	**0.4050**	**0.3184**	**0.3530**
# Like / (# Like + # Dislike)	**0.2320**	**0.2918**	**0.2909**
# Subscribe	0.3028	0.1806	0.2179
# Like / # Day number	0.3140	0.1376	0.1839
# Dislike /# View	0.2072	0.1608	0.1786
# Upload video	0.1946	0.15002	0.1674
# Comment	0.2316	0.1006	0.1336
# View /Length	0.1563	0.0480	0.0724
# Upload video / # Day number	0.1563	0.0521	0.0609
# Dislike / # Day number	0.1195	0.0396	0.0599
# View / # Day number	0.1301	0.0275	0.0524
# Comment / # View	0.1329	0.0245	0.0551
# Subscribe/ # View	-0.0156	-0.1018	-0.0838
Length	-0.0183	-0.10006	-0.0907
# Comment / # Subscribe	-0.1459	-0.1016	-0.1155
# Like / # Subscribe	-0.2026	-0.0986	-0.1272
# Upload video / # Subscribe	-0.1249	-0.2279	-0.2092

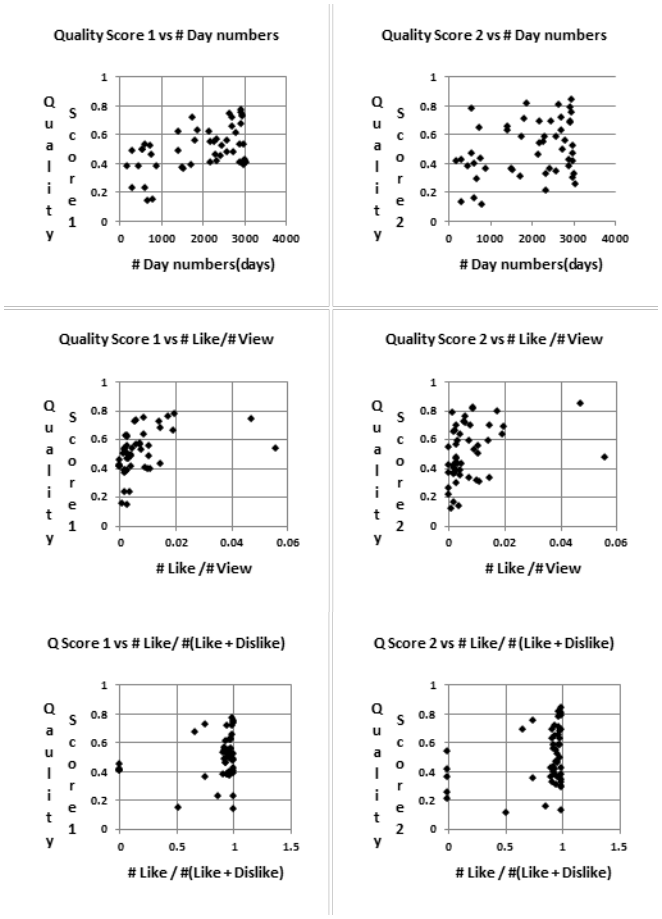

Fig. 1. Scatter plots between the metrics and the quality scores for Group 1 (left panel) and Group 2 (right panel).

Table 3. Performance of the quality estimation model.

	# metrics used	Model parameters $[a_0, a_1, a_2, a_3, a_4]$	Correlation coefficient	RMSE
Group 1	3	[0.304, 2.73x10^{-4}, 2.19x10^{-5}, 2.17, 0]	0.6702	0.110
	4	[0.184, 2.60x10^{-4}, 3.79x10^{-5}, 0.72, 0.13]	0.7071	0.105
Group 2	3	[0.334, 2.75x10^{-4}, 2.20x10^{-6}, 2.9059, 0]	0.4873	0.165
	4	[0.154, 2.55x10^{-4}, 2.63x10^{-5}, 0.71, 0.19]	0.5554	0.157
Group 1+ 2	3	[0.0.33, 2.74x10^{-4}, 7.83x10^{-6}, 2.74, 0]	0.5485	0.143
	4	[0.16, 2.56x10^{-4}, 2.98x10^{-5}, 0.73, 0.17]	0.6108	0.135

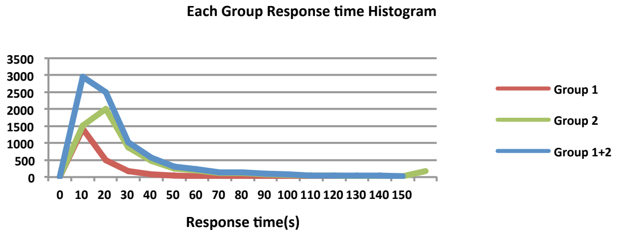

Fig. 2. Histograms of the response time taken for each comparison.

correlation coefficient between the metric and the quality scores obtained from the subjective data. Table 2 shows the results of the correlation analysis, where the ones showing good performance are placed at the top of the table. The ranking of the metrics is slightly different across the groups, but the best four metrics are consistent in Group 1 and Group 2: the maximum size, the day number, the ratio between the like counts and the view counts, and the ratio between the like counts and the sum of the like and dislike counts. The maximum size shows the highest correlation with the subjective scores, which is reasonable because the availability of a high resolution indicates that the source video used for producing lower resolution versions was created at, at least, the high resolution. Then, the day number also shows relatively good performance, as a more recently created video can expected to have better quality in a long term scale. The high ranks of the two ratios of the like counts also make sense in that visual quality is a factor influencing users' satisfaction. Fig. 1 shows the scatter plots between the metrics and subjective quality scores.

4.2. Metadata-based quality estimation model

Based on the above analysis results, we construct a quality estimation model. A linear model is chosen for this purpose, i.e.,

$$\text{Predicted quality score} = a_0 + \sum_{i=1}^{N} a_i m_i$$

where m_i $(i = 1, \dots, N)$ is a metadata-based metric, N is the number of considered metadata metrics, and a_i $(i = 0, 1, \dots, N)$ are the linear model parameters. We considered the top 3 or 4 metrics in Table 2 (i.e., N=3 or 4), and performed linear fitting to optimize the model parameters. The results of the fitting are shown in Table 3 in terms of the linear correlation coefficient and the root mean square error (RMSE) between the ground truth subjective scores and the predicted scores. The best performance is obtained when N=4 for Group 1; the correlation between the predicted quality scores and subjective quality scores is as high as about 0.7, and the RMSE is about 0.1, which corresponds to only 10% of the whole range of the quality scores.

4.3. Comparison of Group 1 and Group 2

Overall, the performance of the metadata in quality score prediction is better for Group 1 than Group 2. Reliability of the subjective data of the two groups is different due to the different levels of 'controllability', i.e., it can be said that Group 1 was more controlled during the experiment than Group 2 and thus produced more reliable data.

In Fig. 2, we compare the histograms of the response time for the two groups. It is observed that the subjects in Group 2 spent more time than those in Group 1. On average, the subjects in Group 1 spent 25.8 seconds per comparison, while 37.5 seconds were spent per comparison for Group 2. Several factors may be involved in producing this difference. The test instruction was directly given to the subjects in Group 1, so it is possible that they understood well the procedure and were able to conduct the experiment efficiently within shorter durations. Moreover, they were mostly young university students in 20s, possibly familiar to watching YouTube videos, while the subjects in Group 2 would be more diverse in their background and age.

Although taking more time in watching videos may mean being more careful in judgment, it may also have an undesirable influence on the results; as a subject watches more, he/she may be attended more to the contents and thus the final decision may be biased due to the preference of the contents.

5. CONCLUSION

In this paper, we explored the feasibility of the metadata as the indicator of quality of online videos. Crowdsourcing-based subjective quality assessment was conducted to collect ground truth subjective quality data, which were used to evaluate various metrics derived from metadata for quality score prediction. It was found that the maximum video resolution, the day number of the uploaded date, and the relative like counts were effective. Then, quality estimation models were constructed as a linear combination of the well-performing metadata metrics, which was shown to be successful by showing correlation up to about 0.7.

Our results are promising in the sense that the metadata alone can already show such good performance and, thus, synergy is expected when they are combined with no-reference quality metrics analyzing visual data, which will be pursued in our future work.

5. REFERENCES

[1] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 660-668, 2008.

[2] T. Xia, T. Mei, G. Hua, Y.-D. Zhang, and X.-S. Hua, "Visual quality assessment for web videos," *Journal of Visual Communication and Image Representation*, vol. 21, pp. 826-837, 2010.

[3] C.-C. Wu, K.-T. Chen, Y.-C. Chang, and C.-L. Lei, "Crowdsourcing multimedia QoE evaluation: A trusted framework," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1121-1137, 2013.

[4] Recommendation ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," 2012.

[5] J.-S. Lee, "Paired comparison for subjective multimedia quality assessment: Theory and practice," in *Proc. Int. Symp. Circuits and Systems*, May 2013, pp. 1099-1102.

[6] J.-S. Lee, "On designing paired comparison experiments for subjective multimedia quality assessment," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 564-571, 2014.

[7] S. Hacker and L. von Ahn, "Matchin: Eliciting user preferences with an online game," in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, Apr. 2009, pp. 1207-1216.